**Airbnb Price Analysis**

Yuan Shi, Xin Tang, Sichen Tang

Dec/12/2016

**1. Introduction**

New York City is the second largest city around the world providing Airbnb housing. So NYC Airbnb is chosen as the object of our project. The purpose of the project is to get some general insight from history Airbnb housing data, and predict the price of Airbnb housing based on price data of all the Airbnb housing in NYC, with the regression tools we learn this semester. The results of the project could be used as guideline for Airbnb hosts to get an overview of the market. And our model suggests them a suitable price for their own housing. Also, it could be a decide model for guests as well, to check whether the housing provides an appropriate price.

Data we used:

1. New York City’s Airbnb data from 2015 (missed February and July) and July 2016.

<http://insideairbnb.com/get-the-data.html>

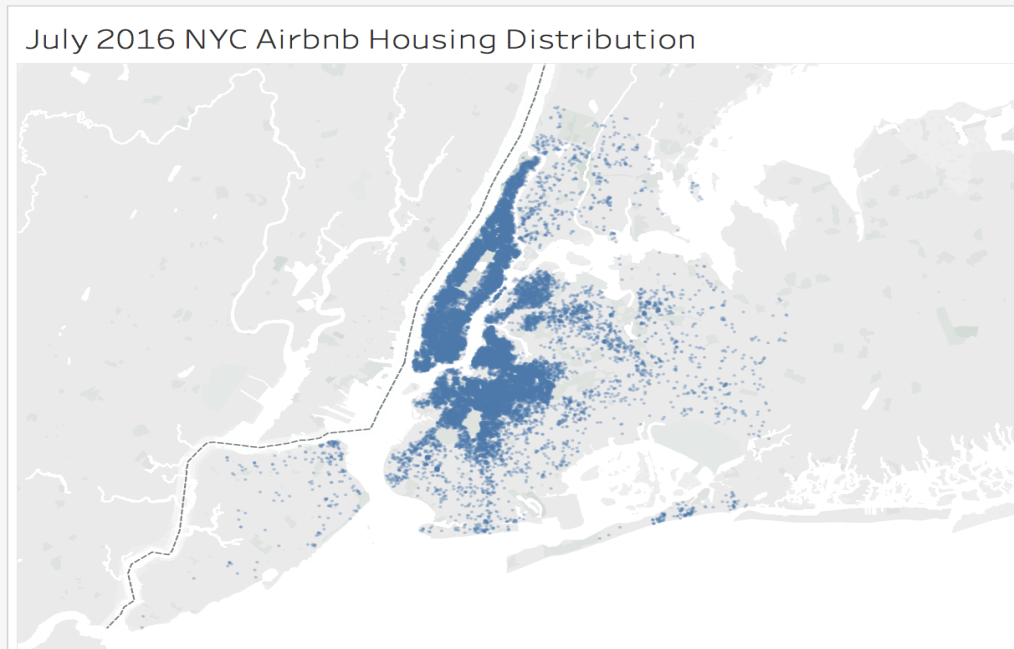
As Airbnb hasn’t opened data through API to public, the data we get is through an independent group, who crawled data from Airbnb website from time to time.

2. NYC zip code and borough name data: <https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm>

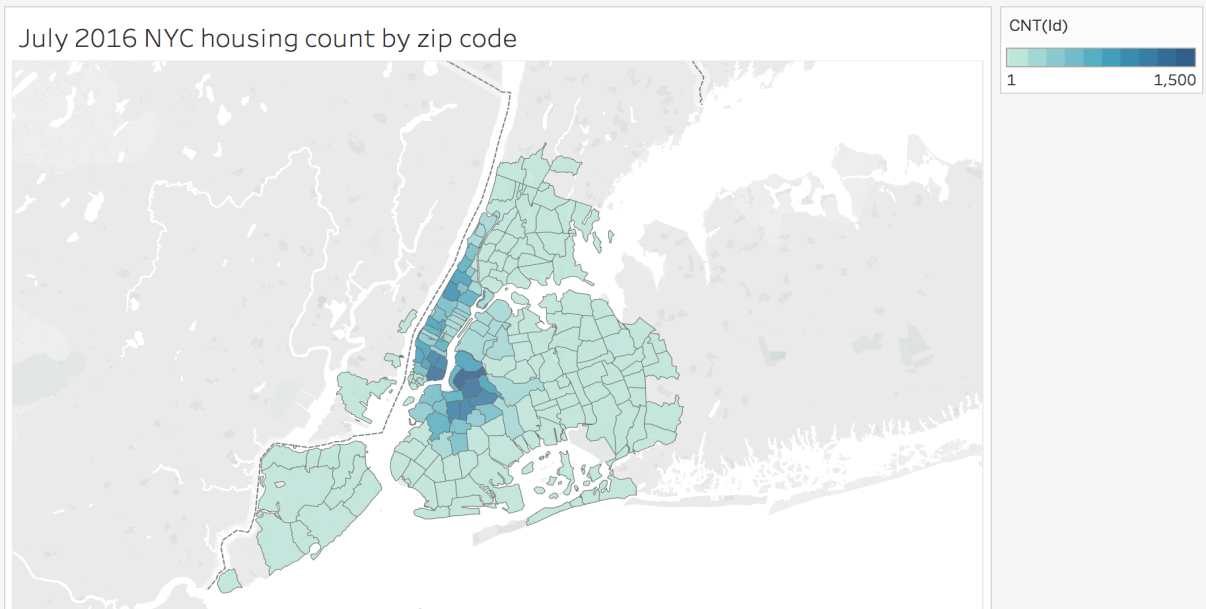
**2.General information**

1.Where are the housings?

All the Airbnb housings are plotted, based on the latitude and longitude information in 2016 July dataset. As the location information is anonymized by Airbnb, so the location information in our data will be 0-450 feet (150 meters) off the actual address. Therefore, the point in our map is not at the exact location where the housing is, but 150-meter error. At NYC scale, this is bearable, and still reflects where the housing is supplied more or less. The housings in the same building are anonymized by Airbnb as well, they will be scattered around the actual address. That is, no two points in map are totally overlapped.



There are 38810 housings mapped. It is very clear that entire Manhattan has a great deal of Airbnb supply. But as the data point is too much, looks like they are equally distributed. Housing count by zip code could more clearly show. The minimum count is 1, and the maximum count is 1953. Except 1953, the other are all below 1500. So group counts from 1 to 1500 into 10 groups.



There are three dark blue areas. First is around Williamsburg, where young people gathered and would be more likely to open their doors to strangers, which makes sense. Second and third are downtown and upper west Manhattan.

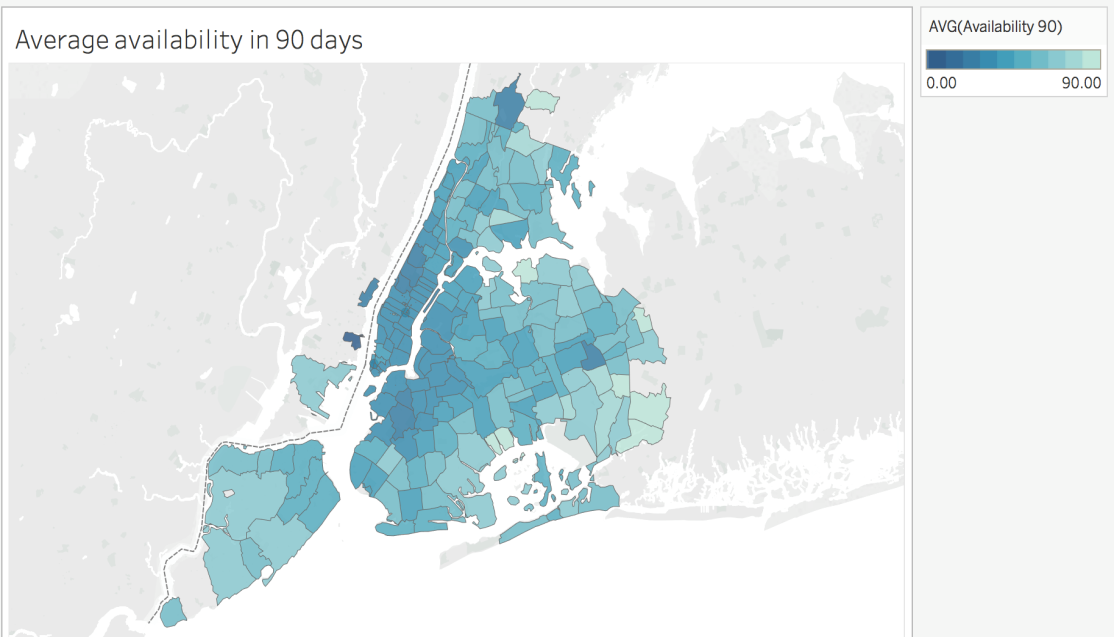
2. Where is cheap and where is expensive?



Calculate average price of housing within each zip code. The lowest average price is $45, the highest is $449. Except for $449, the other average prices are below $350. So group the average price into ten groups from $45 to $350. Basically the most expensive area is Manhattan. From downtown to uptown, the average prices of each zip code are above $160. Besides, average price of zip code 11365 (Queens), 11374 (Queens), 10302(Staten Island) are very high. Checked their housing count, 11365 has only 2 housings; 11374 has 56 housings; 10302 has 5 housings. So the price could be strongly influenced if there are some very expensive housings. As we compare price of all the house types and room types together, place as Staten Island and Queens are more likely to supply villas or big apartments to guests, which bring up average price.



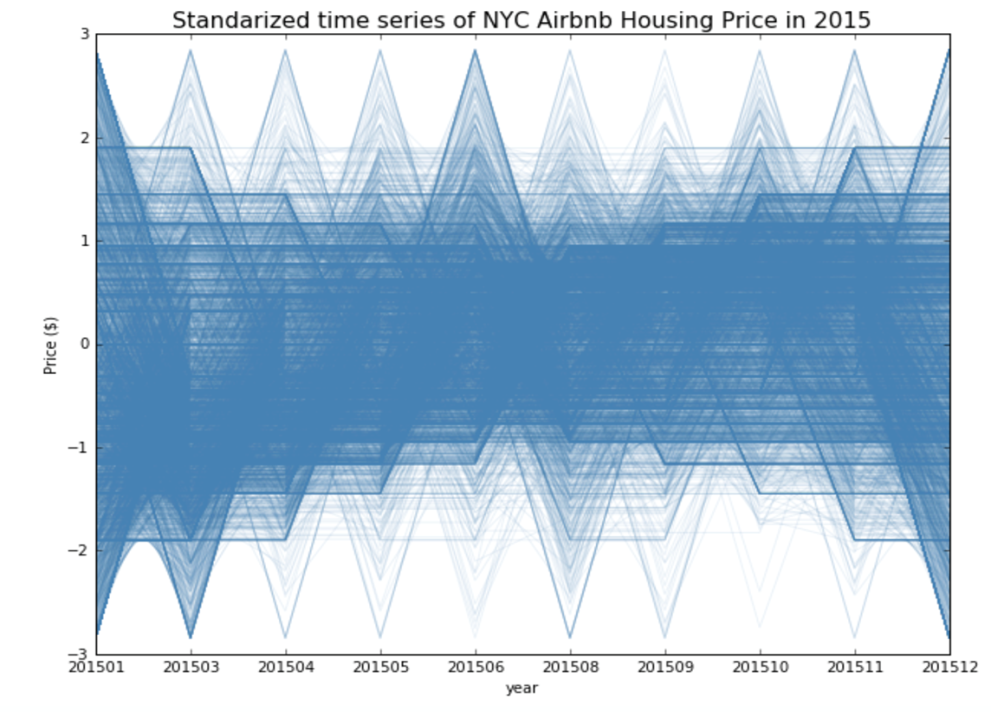
3. Where are the popular place guest chose?



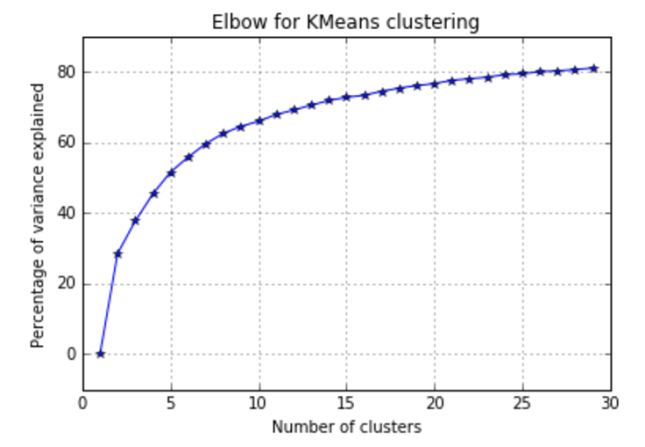
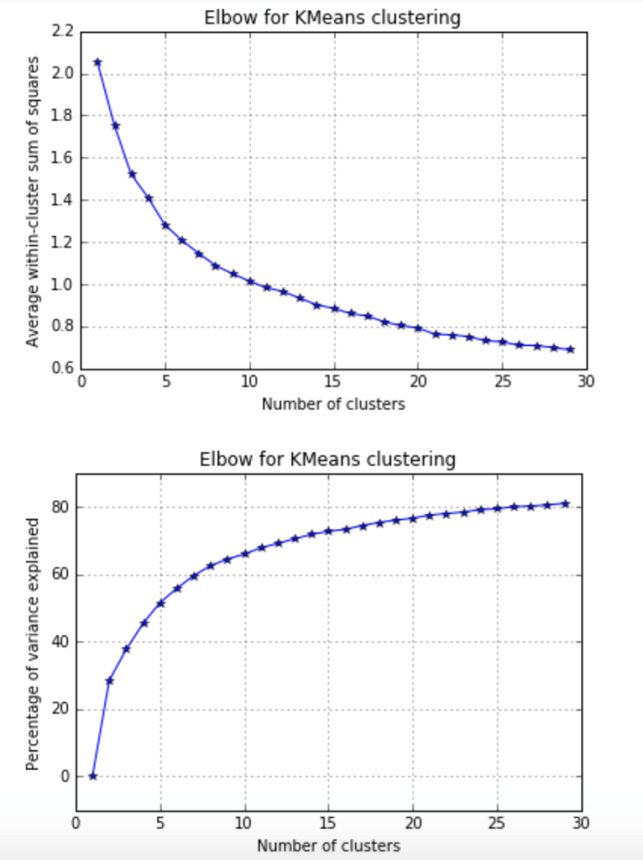
Availability in 30 days means that at the time point the data is collected ( July 2nd, 2016) , how many days between July 2nd to August 1st are left to be ordered. This is not occupancy rate. Because the room now is available probably get taken later. But this availability number somehow shows an area is popular or not. If an area is popular, they will be booked in advance. Not surprisingly, Manhattan and near-Manhattan Brooklyn are more popular. Mostly availability is less than 10 days.

4. What is housing price trend in 2015?

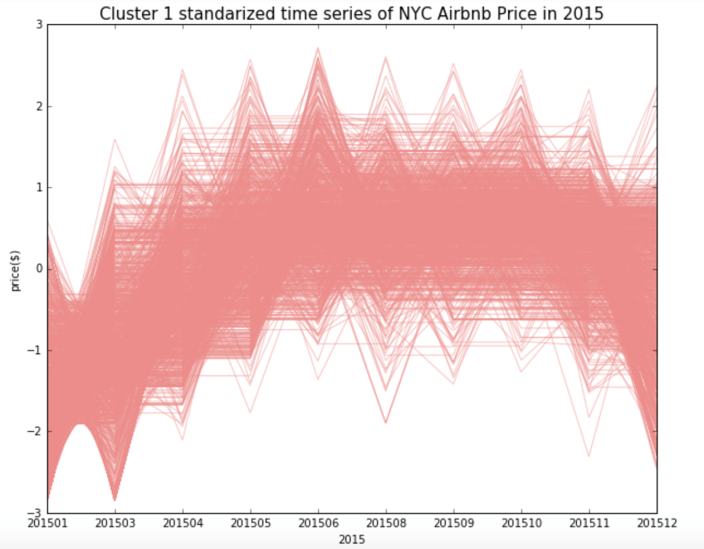
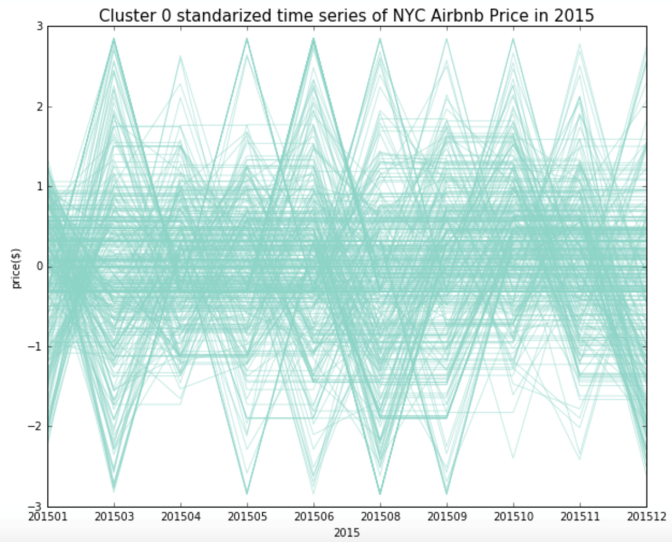
Months except Feb and July in 2015 have more than 30000 housing information. Merge the price data on Airbnb ID, if a housing has Nan price in some months, it will be dropped. After cleaning, the price data of 2015 has 8837 housing information. Standardize price to compare their trend, rather than the absolute value of the price.

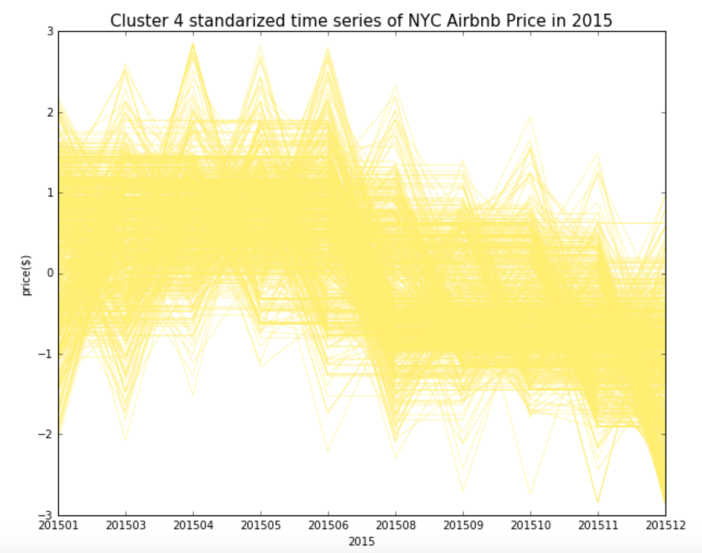
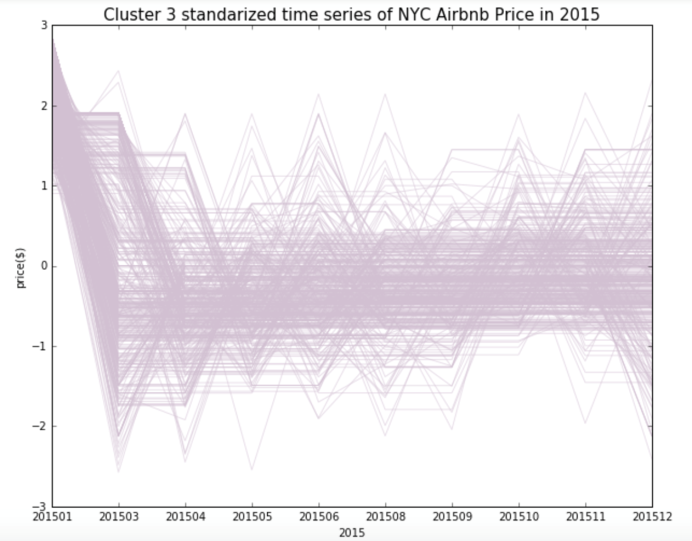
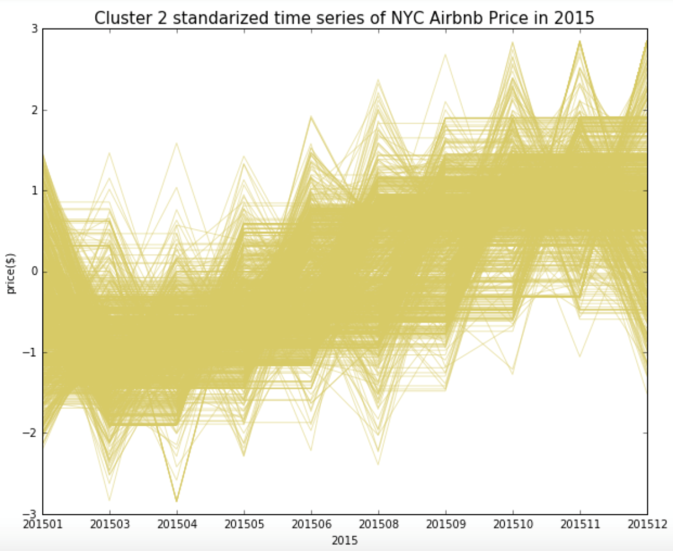


Then use elbow for K-Means clustering to find out suitable number of clusters.



The elbow line is really smooth, chose K= 5 to try. From sklearn.cluster import Kmeans to cluster the standardized price time series.





Cluster 0 has 3503 housings, 39.6% of all. The trend is fluctuating around a certain price.

Cluster 1 has 1631 housings, 18.5% of all. The trend is rising during summer, and going down in winter.

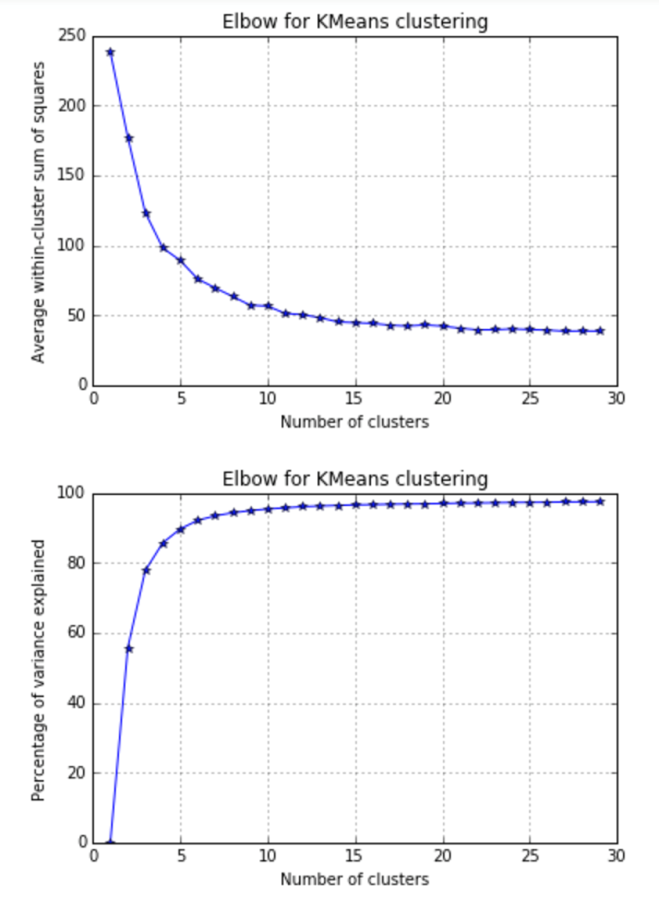
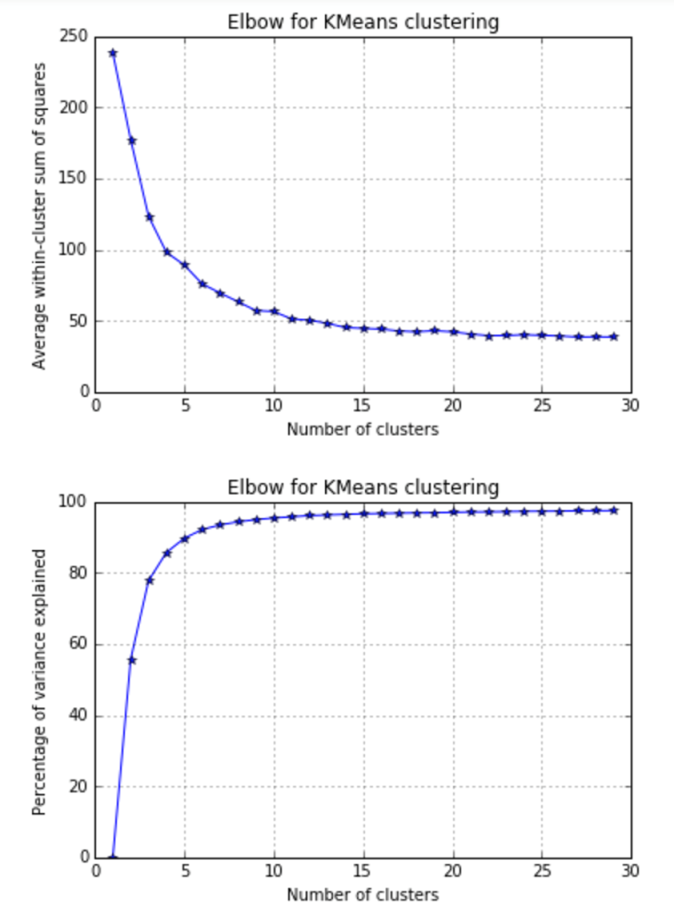
Cluster 2 has 1865 housings, 21.1% of all. The trend is going down in Spring and rising in Fall.

Cluster 3 has 725 housings, 8.2% of all. The trend is going down in Spring and almost be stable afterward.

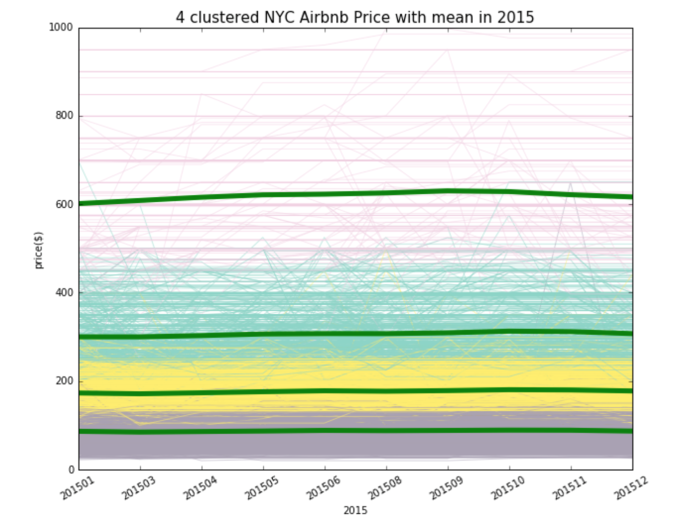
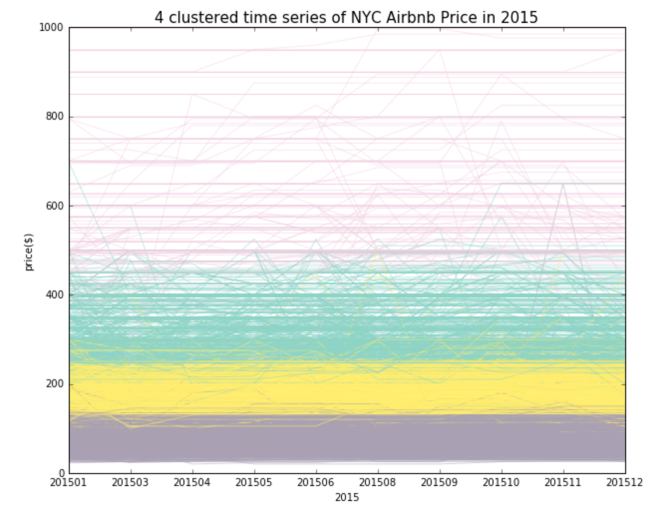
Cluster 4 has 1113 housings, 12.6% of all. The trend is stable first half year and going down the last half year.

5. How is different level of housing?

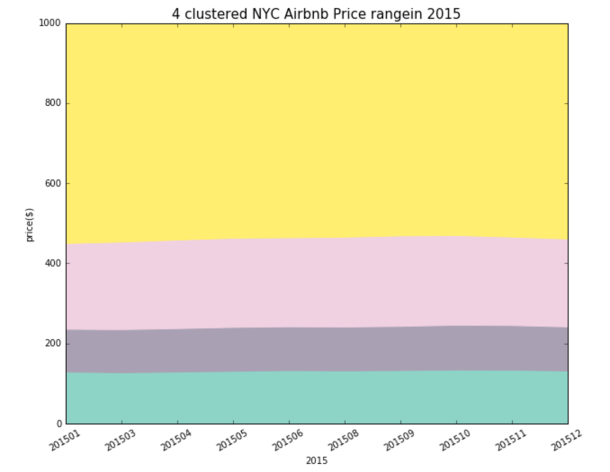
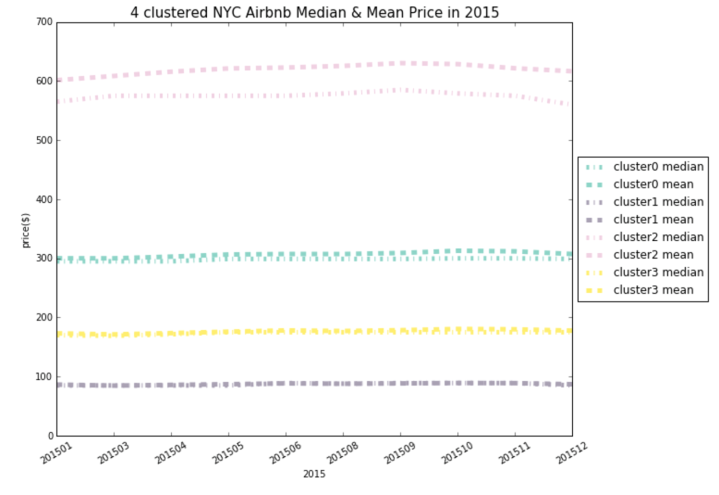
Use K-Means to cluster the non-standardized price. First, decide how many levels would best describe the price? Use Elbow to choose K=4.



Cluster all the housing price. Green lines show the mean price of each group.

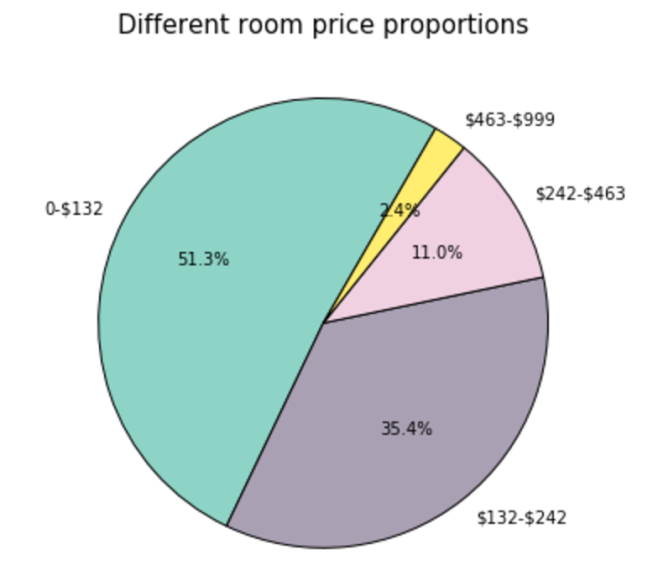


The pink group, the most expensive housing, has widest range. And two cheapest groups are more dense. So is it suitable to use mean price to describe the group?



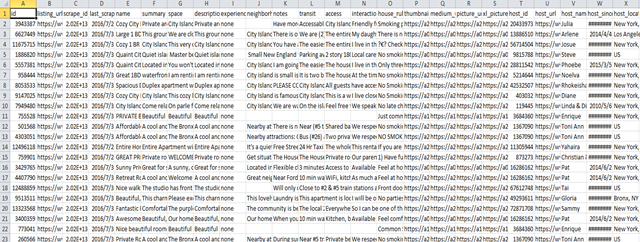
The above left plots mean and median of each group. Except the most expensive group, at the other three groups, mean and median is overlapped. So in those three groups, mean could well present the group. At the most expensive group, median is much lower than mean, that is, the group mean is pushed higher by some extreme expensive housings. So in the most expensive group, mean is probably not representative.

The above right shows each group's range. 0- $132 ; $132- $242; $242-$463; $463-$999 are the four groups. Each group's proportion shows as below. Housing between 0-$132 still is the majority.



**3.Linear Regression**

3.1 Original Data (<http://insideairbnb.com/get-the-data.html>)



3.2 Variables

From the original data source we select the useful features below:

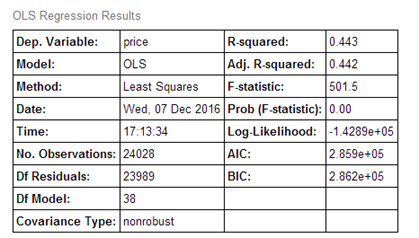
Price/ host response rate/ host has profile picture/ host identity verified/zip code

/property type/room type/accommodates/bathrooms/ bedrooms/ beds/ bed type

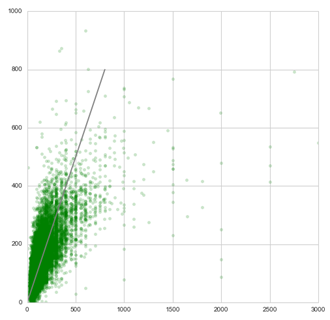
/extra people/minimum nights/review scores rating/cancellation policy

3.3.Model

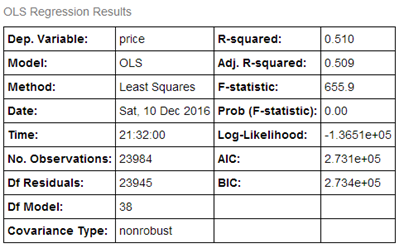
Then we used the statsmodels.OLS to fit the model. We define price as the dependent variable, other above features as the independent variables. For the string variables (such as cancellation policy and room type) we use the dummies to label them as numbers. We also include zip codes as geology variables. (Manhattan=0, Brooklyn =1, Bronx=2, Queens=3, Staten=4)

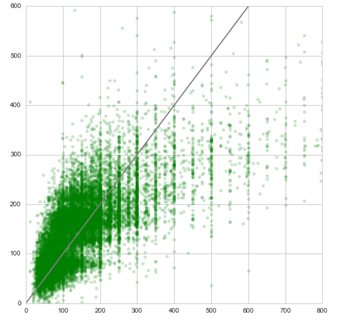


As the OLS Regression Results shown, the R-squared is 0.443.



The quantity of price > 1000 is very small, we exclude that price and fit the model again.





The line is 45 degree. We can see that the predicted price and the actual price fit in a somewhat reasonable range. Also, the R-squared as shown above changes to0.510, which becomes larger than the previous R-squared.

3.4 Some significant variables

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | Standard deviation | P-value |
| Accommodates | 12.0220 | 0.522 | 0.000 |
| Bathrooms | 50.4762 | 1.427 | 0.000 |
| Bedrooms | 30.9278 | 0.970 | 0.000 |
| Review scores rating | 0.9072 | 0.056 | 0.000 |
| Loft | 40.6433 | 9.796 | 0.000 |
| Entire home/apt | 48.3300 | 4.046 | 0.000 |

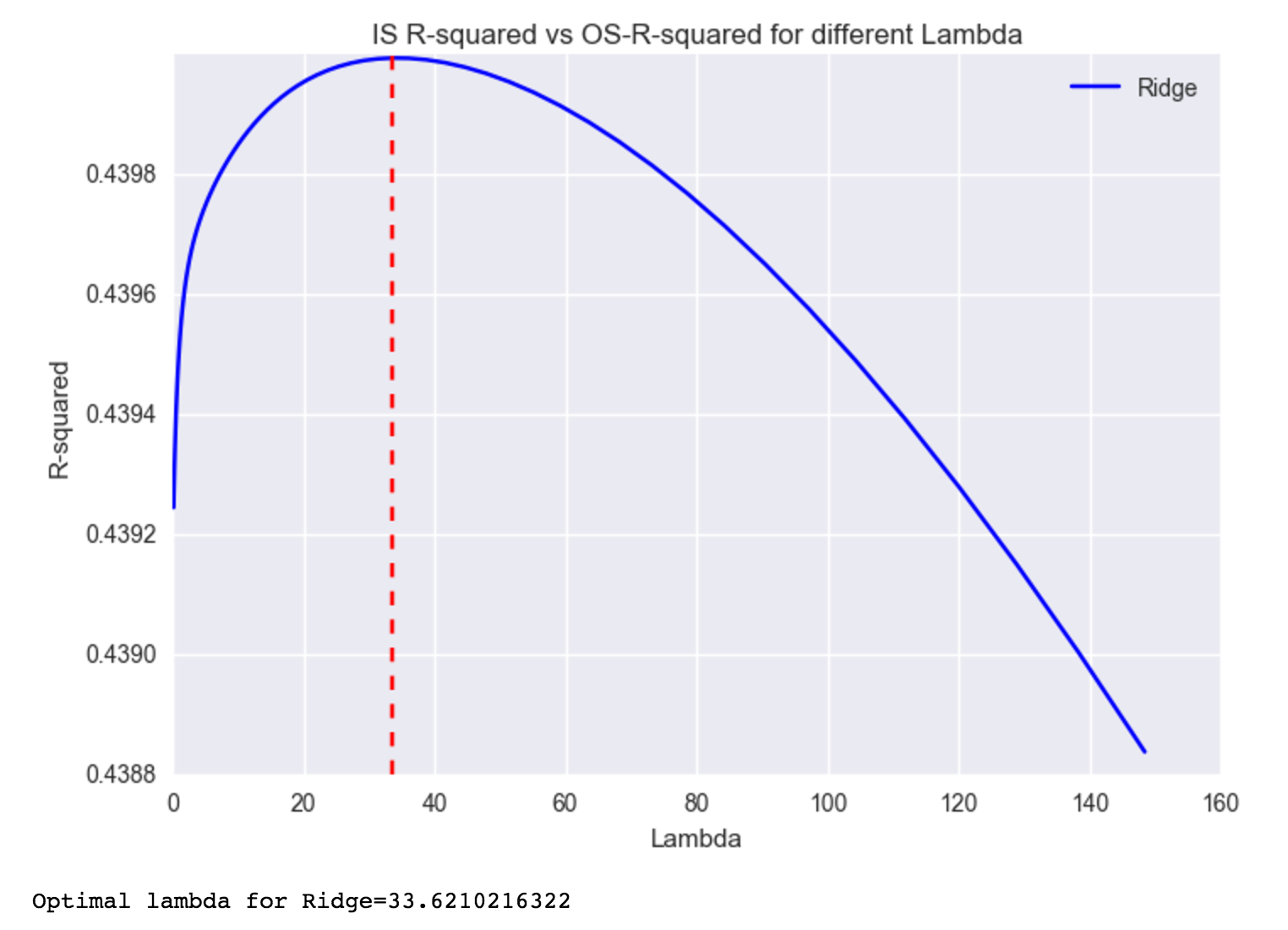
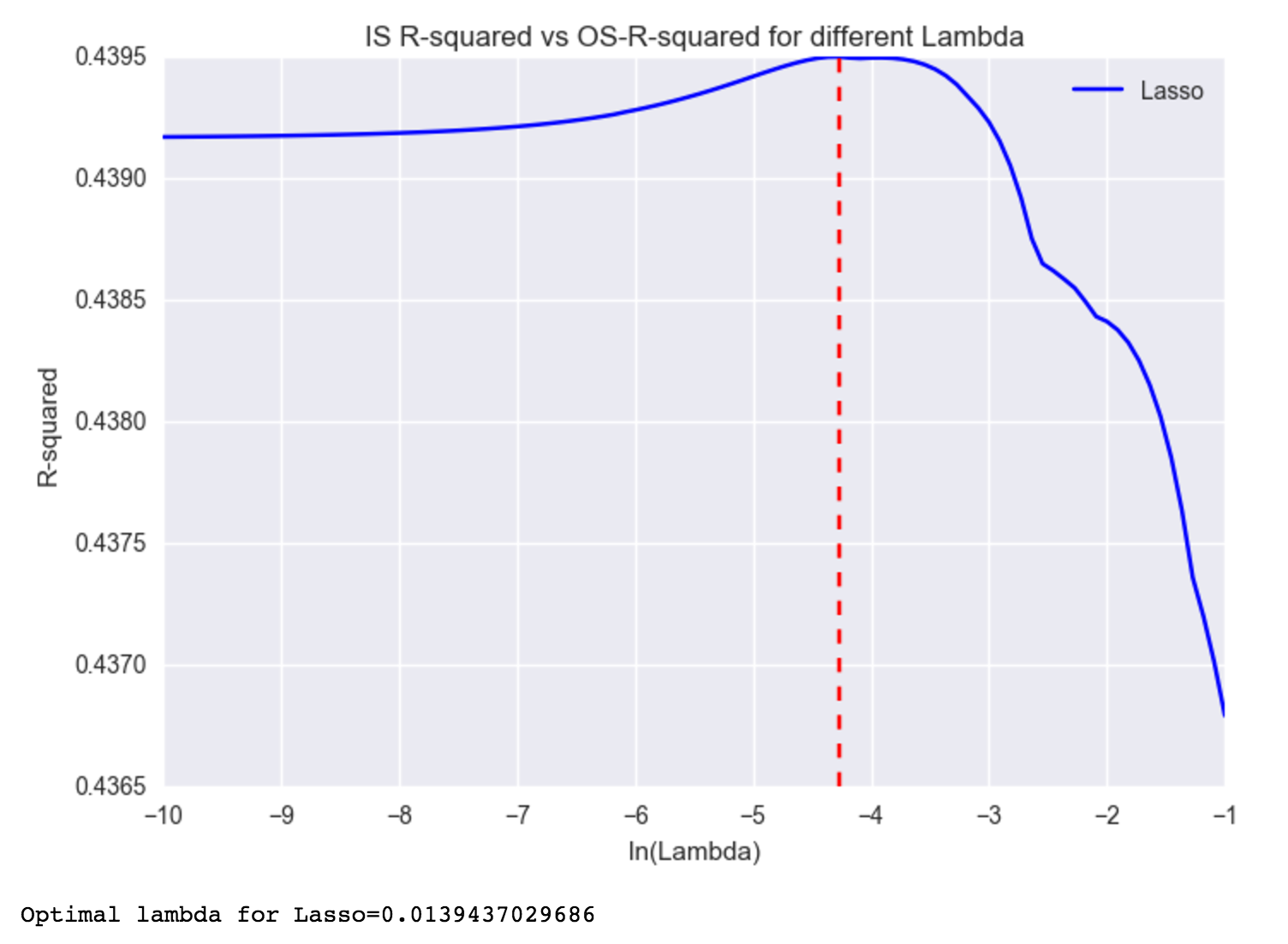
As shown above, we select some significant variables whose p-value is 0.000 < 0.05. Accommodates has a positive coefficient 12.0220, that is, while the housing can accommodate more people, its price will be higher. Similarly, while the housing has more bathrooms, or more bedrooms, higher review scores rating, its price will increase. Besides, on the one hand, if the housing is a loft style, it will also have a higher price. On the other hand, if the room type is entire home or apt, the housing will get a high price.

**4.Lasso & Ridge regression**

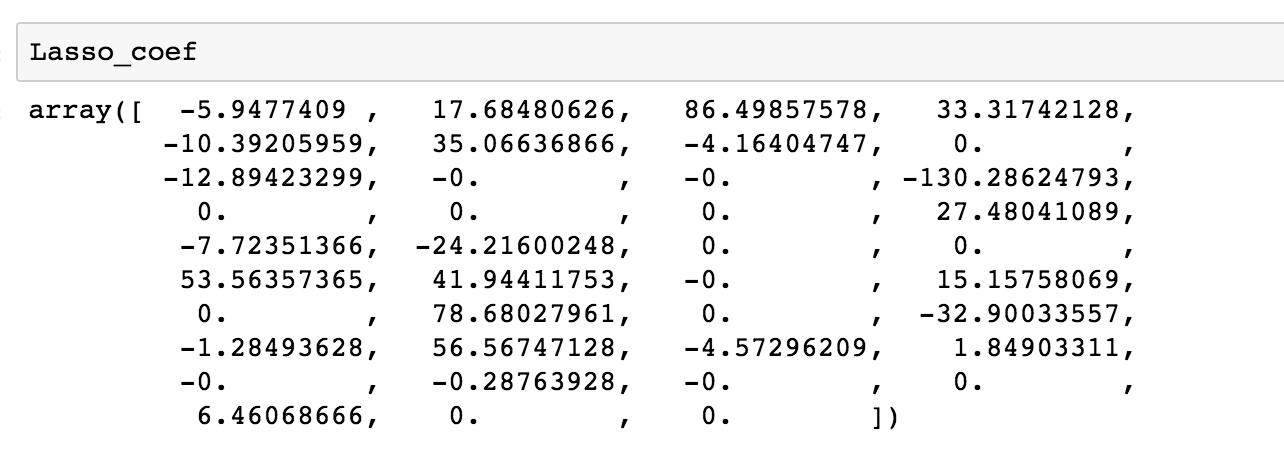
We fit a model of price across to the features we selected, we deleted features those contain natural language, since we are not able to extract the information into a type that can be used for a linear model, though we believe that the information they contained may be correlated to price.

From the summary of the linear regression we find that the r square is around 0.42. So we are thinking other than to select the features by ourselves, it may be better for us to select the features by Lasso regression.

First, I tried to find the appropriate lambda for Lasso regression and ridge regression.



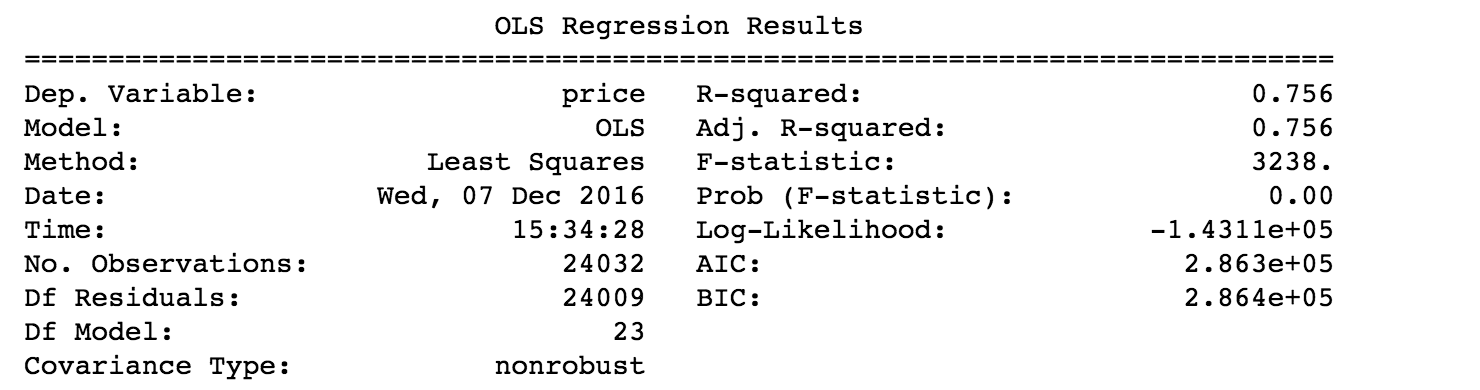
And then I fit the data with lasso regression, from the lasso coefficient we find that there are still some columns can be deleted in the linear model.



And then we keep the columns with non 0 coefficient.



Then I deleted the columns with coefficient 0. And fit the data again. We have a much greater r square which means that this model explains the relationship between other features and price better.



Also We have tried to include the borough variable in our model, we clustered the borough by zip code, and generated it as a dummy variable. After we fit the model with this information the linear regression does not yield too much difference in accuracy if we check the r square. But when I use the lasso regression, I find that I only have 4 coefficients is 0. It tells us that the borough dummy variable confused the lasso regression.

**5.Method Conclusion & Possible improvements in future**

1.Method conclusion

(1)Time series

(2)Cluster:K-means

(3)OLS regression

(4)Lasso and Ridge

2.Improvements

(1) Count housings by zip code could show where is the zip code with more housing. But sometimes the zip code area is larger, and then it covers more housing supply. So if we can get the data of area each zip code, and calculate housing count per square meters by zip code, the heat map would make more sense.

(2) In this project we basically working on numeric variables, those are easily used in fitting a linear model. But we know that the data include more useful information can be used to predict the price, such as the customer review and regional information. How to extract important information from natural language will be our future work. We tried to find some keywords in reviews so that we can use the frequency as the test feature, but it doesn't work out. We plan to put more effort in this part in the future.